Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 106 (2013) 394 – 400

Procedia
Social and Behavioral Sciences

4th International Conference on New Horizons in Education

An investigation of goodness of model data fit

Gülşen Taşdelen Teker[a*] Hülya Kelecioğlu[b] Melek Gülşah Eroğlu[c]

[a]*Sakarya University, Faculty of Education, 54300, Sakarya, Turkey*
[b]*Hacettepe University, Graduate School of Social Sciences, 06800, Ankara, Turkey*
[c]*Gazi University, Faculty of Education, 06500, Ankara, Turkey*

Abstract

IRT models have many advantages over CTT. However, the model-data fit should be verified as a prerequisite to use IRT models. Therefore, in this study it is aimed to investigate which IRT model will provide the best fit to the data obtained from SBS 2009 science subtest. For goodness-of-fit analysis, first the model assumptions and then the expected model features were tested. The model assumptions unidimensionality and local independence were investigated. In the expected model features part the invariance of ability parameter estimates and invariance of item parameter estimates were analyzed. To determine the best model, the results of the chi-square statistics of -2 log likelihood values of models were compared. The results suggested that three parameter logistic model is the most appropriate model for data fit.

Keywords: Item response theory, model-data fit analysis, person and item statistics

1. INTRODUCTION

The science of psychological tests is known as psychometrics. Psychometrics had its origins in the mid to late 1800's as part of an effort to quantify cognitive and affective attributes in humans (Crocker & Algina, 1986). Psychological tests scores are calculated, they can be used to compare achievement results at the student, school or state level. The scores then analyzed in a way that their results can be appropriately interpreted. The classical test theory and item response theory are used in practice. In classical test theory, the model of measurement error is based on the correlation coefficient. The correlation coefficient, developed by Charles Spearman, attempts to explain error using two components: a true correlation and an observed correlation (Crocker & Algina, 1986).

* Corresponding author. Tel.: +90-264-614-1033.
E-mail address: gtasdelen@sakarya.edu.tr

doi:10.1016/j.sbspro.2013.12.045

Mathematically, this can be written as Xp = Tp + Ep where Xp is the observed score, Tp is true score and Ep is the error associated with the score. The observed score is the actual score an individual received on that particular test and the true score is the average score an individual would receive on a test if they repeated it many times. Unfortunately, classical test theory is limited to simply describing the total test score for one particular test and group of examinees.

The initial theoretical groundwork for an alternative to the CTT paradigm known as Item Response Theory (IRT) was laid down in the 1940's by D. N. Lawley (McDonald, 1999). A more comprehensive theoretical framework for IRT was subsequently developed and shaped from the 1950's to the 1970's by a number of prominent psychometricians and statisticians, including Frederick Lord, Georg Rasch and Benjamin Wright (Baker, 2001; McDonald, 1999). The mathematics and computational algorithms of IRT are far less tractable than CTT and therefore the practical implementation of IRT had to wait until the 1970's and 1980's for the development of computers that could carry out the necessary computations for IRT.

Item response theory differs from classical test theory by focusing on measuring some latent trait possessed by the examinees, usually ability. The key feature of IRT that sets it apart from CTT is the explicit mathematical modeling of the stochastic relationship between the performance on an individual item and the underlying continuous scale of the latent construct that the item is theorized to be measuring (Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991). IRT uses a mathematical model to express the probability of an examinee answering an item correctly as a function of the examinee's ability. These functions (for each item) are the item response functions (IRFs) which serves as the mathematical model linking the observable data (item performance) to the unobservable data (ability). One useful feature is that of the test characteristic function that is the sum of the item characteristic functions that makes up a test and can be used to predict the scores of examinees at given ability levels. If the test is made up of test items that are relatively difficult, then the test characteristic function is shifted to the right and examinees tend to have lower expected scores on the test than if easier test items are included.

Unlike classical test theory models, the model parameters in IRT can be adjusted to account for different ability distributions possessed by the examinees. This is known as the invariance property. That is item parameters estimated using one group of examinees can be transformed to match those from another group of examinees. If the item is binary (the item is either correct or incorrect) then the IRFs are commonly assumed to be monotonic, unidimensional, and locally independent. Monotonicity means that as the examinee's ability increases, the probability that the examinee will respond correctly (for a particular item) also increases. Unidimensionality means that only one ability or trait is necessary to explain for examinees' test performance. Item response models that assume a single latent ability are referred to as unidimensional (Hambleton & Swaminathan, 1985). The assumption cannot be strictly met since there are always other factors affecting the test performance. Therefore, the requirement for the assumption is to have a dominant factor. Local item independency means that an examinee's responses to different items in a test are statistically independent. For this assumption to be true, an examinee's performance on one item must not affect, either for better or for worse, his or her responses to any other items in the test (Hambleton & Swaminathan, 1985).

The item analysis in the item response theory consists of (a) determining sample-invariant item parameters using relatively complex mathematical techniques and large sample sizes, and (b) utilizing goodness- of- fit criteria to detect items that do not fit the specified response model. The property of sample invariance inherent within IRT means that test developers do not need a representative sample of the examinee population to calibrate test items. They do, however, need a heterogeneous and large examinee sample to insure proper item parameter estimation. Also IRT requires larger sample sizes to obtain good item parameter estimates, the test

developer must ensure that the examinee sample is of sufficient size to guarantee accurate item calibration (Hambleton& Jones, 2005).

However, measurement specialists cannot benefit from the advantages mentioned above unless model data fit is achieved (Hambleton, Swaminathan, & Rogers, 1991). Although there are several studies that investigated model-data fit, few studies investigated the fit of IRT models to data obtained from achievement tests administered to elementary school students. Therefore, in this study model-data fit investigations were conducted on the data obtained from examinees that were preparing for Seviye Belirleme Sınavı (SBS) in Turkey.

2. METHOD

2.1. Study Group

This study examined SBS science subtest data for 2009. SBS is administered to 8th grade students. Total of 1 million 15 thousands of students took the 2009 SBS and all of them were responsible for answering questions in science subtest constitute the population of the study. Among the students took the exam, 1964 students were randomly selected as the study group. Among 1964 subjects, 56.2 % of them is male and 43.8 % of them is female.

2.2. Instrument

The data set was obtained from the 8th grade examinees that took the SBS 2009. SBS composed of four subtests which are Turkish, Science, mathematics and social sciences. In this study the science subtest of the SBS was used. The science subtest consist of 20 items. However, the result of factor analysis showed that to make it unidimensional, 5 of the items were deleted and the analysis was done with 15 items.

2.3. Design of the Study

Methods for assessing goodness of fit were presented by Hambleton et al. (1991). Checking model assumptions and checking expected model features were two important goodness of fit investigations. In the first part of the analysis in which model assumptions held was investigated through analysis of unidimensionality and local independence. In the second part, degree to which desired model features were obtained was investigated through analysis of invariance of item parameter estimates and invariance of ability parameter estimates.

2.4. Assessment of Goodness of Model-Data Fit

Item response models offer a number of advantages for test score interpretations and reporting of test results. However, the advantages will be obtained in practice only when there is a close match between the model selected for use and the test data. This can be examined through two step processes: checking the model assumptions and model features.

2.4.1. Checking Model Assumptions

There are two main assumptions which are unidimensionality and local item dependence.

Unidimensionality: To check the unidimensionality assumption factor analysis was conducted. Eigenvalues and obtained scree plot were investigated in order to determine whether there was a dominant first factor. According

to Hambleton, Swaminathan, and Rogers (1991) a dominant first factor is needed to satisfy unidimensionality assumption. In other words, there should be a large difference between the first eigenvalue and second eigenvalue. Moreover, a significant drop in the contribution of the factors between the first and second factors can be seen as an evidence for unidimensionality. To do the factor analysis to the binary science items STATISTICA computer program was used to get the tetra choric correlation matrices.

Local Item Independence: According to Hambleton Swaminathan, and Rogers (1991) when unidimensionality assumption is met the local independence assumption is also satisfied. Therefore, the clues for the unidimensionality was used also for the local item independence.

2.4.2. Checking Model Features

Invariance of Item Statistics: Invariance of item statistics can be defined as item statistics are obtained that do not depend on the sample of examinees used in the calibration of test items. To investigate the degree to which the property of invariance held for the item difficulty and item discrimination parameter estimates, the difference is that extreme ability groups (e.g. random groups, high vs. low ability groups) were formed and item parameter estimates in the two samples were compared.

Invariance of Ability Estimates: Invariance of ability statistics can be defined as examinee ability estimates are obtained on the same ability scale and can be compared even though examines may have taken different sets of test items from the pool of test items measuring the ability of interest (Hambleton & Swaminathan, 1985). To investigate the degree to which the property of invariance held for the ability parameter "θ" estimates, by administering examinees two or more samples of test items that vary widely by means of some parameters (e.g. difficult vs. easy, even vs. odd items).

To do the IRT analysis to check the invariance of item and ability statistics BILOG computer program was used.

3. RESULTS

3.1. Checking Model Assumptions

To test the unidimensionality assumption of IRT, factor analysis was conducted with 20 science items. The results showed that the test did not met the assumption of unidimensionality. Therefore, five items were deleted and the final results were satisfactory. Both the initial and the final factor analysis results are at Table 1.

Table 1. Total Variance Explained

Component	Initial Eigenvalues			Final Eigenvalues		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7.31857	36.59285	36.59285	6.728082	44.85388	44.85388
2	1.55727	7.78633	44.37918			

As stated before, local item independence is met with the satisfaction of unidimensionality (Hambleton, Swaminathan, & Rogers, 1991). Therefore, no extra analysis were done to check local item independence assumption.

3.2. Checking Expected Model Features

Invariance can be thought as both item and ability levels. If a test is scaled with IRT, then it can be used many times to different examinees without any change in item parameters and examinees' ability can be measured although by taking different items. To examine the invariance of item and ability statistics, all the items were scaled by means of three item response models which are 1-, 2-, and 3-PL models. The summary parameters are in Table 2.

Table 2. The summary of Item Parameters Scale Values

Model	Parameter	Mean	Sd	Min.	Max.
1PL	b	0.141	0.524	-0.757	1.069
2PL	a	1.396	0.379	0.777	1.883
	b	0.184	0.605	-0.658	1.505
3PL	a	2.547	0.783	1.481	3.972
	b	0.514	0.510	-0.403	1.412
	c	0.173	0.035	0.118	0.234

Invariance of Item Statistics: To investigate the degree to which the property of invariance held for item parameters under each model, correlation analysis was conducted on samples (random group and high-low ability group). The results are in Table 3.

Table 3. The Correlation Coefficients of Item Parameters from Different Ability Groups

Model	Parameter	Groups	
		Random groups	High-Low Ability
1PLM	b	0.991	-
2PLM	a	0.934	-0.322
	b	0.988	0.178
3PLM	a	0.788	-
	b	0.986	-
	c	0.705	-

Before the interpretation of the results of invariance of item statistics, it can be seen from the Table 3 that there is no values for the high-low ability group for 1- and 3-PL models. Since b parameter for 1PLM and a,b and c parameters for the 3-PLM for low ability group could not estimate through the analysis done by BILOG, their correlations with the high ability group could not also estimated. There is another point can be seen from the Table 3 that the estimated correlations for the 2-PLM for high-low ability groups were very low. Therefore, only the results of random groups were used for the evidence of invariance of item statistics.

All results obtained from random sample group for the item difficulty parameter "b" showed very strong correlation coefficients (the lowest 0.986 and the highest 0.991). These strong correlation coefficients are excellent indicators of invariance.

Compared to other IRT models, correlations under 1- and 2-PLM were quite strong; therefore invariance property was best achieved under 1- and 2-PLM. Moreover, among the three models, the lowest coefficients were obtained under 3PL model but it is still very high. Invariance of discrimination parameter "a" was also investigated on random group samples. Correlations under 2-PLM was much stronger compared to 3-PLM. In addition, correlations obtained for invariance property of discrimination parameter were weak compared to correlations obtained for item difficulty parameter. Moreover, as the variability in sample increased the correlation

coefficients obtained for invariance property of both item difficulty and item discrimination parameters decreased. Compared to 3-PLM, 2-PLM provided well fit when invariance property of discrimination parameter is considered.

Invariance of Ability Parameter Estimates: To investigate the degree to which the property of invariance held for the ability parameter estimated under each model person statistics obtained on samples (difficult-easy and odd-even samples) was correlated. For all the IRT models the correlations of ability estimates on odd-even sample was high enough for 1- and 2PL models (see Table 4). However, for 3PLM, the coefficient was weak compared to other two IRT models. For all the IRT models the correlations of ability estimates on difficult-easy sample was not very high but it is enough to hold the invariance property.

Table 4. The Correlation Coefficients of Ability Parameters from Different Item Groups

Model	Items	
	Od-Even	Difficult-Easy
1PL	0.715*	0.621*
2PL	0.731*	0.625*
3PL	0.629*	0.649*

*Correlation is significant at the 0.05 level (2-tailed).

3.3. Best Model Fit to Data

To determine which model best fit to the data, the maximum likelihood method was used to get -2Log Likelihood values. These values are in Table 5.

Table 5. -2 Log Likelihood Values for the 1-, 2-, and 3-PLM

Model	-2 Log Likelihood Values
1PLM	34182,0708
2PLM	33874,0544
3PLM	33460,4634

The differences of the values given at Table 5 was important to decide the model appropriate to the data. Therefore the differences between the -2 Log Likelihood values and the decisions were taken and given at Table 6.

Table 6. The Results of Model Differences

Model Differences	Result
(1PLM – 2PLM) = (33874,0544 - 34182,0708) =308,0164 > 24,99579	2PLM is better than 1PLM for model-data fit.
(2PLM – 3PLM) = (34182,0708 - 33460,4634) = 413,591 > 24,99579	3PLM is better than 2PLM for model-data fit.

As a result, we can conclude that, for the model investigated, 3 PLM is best fit.

4. DISCUSSION AND CONCLUSION

The purpose of this study was to investigate which IRT model would provide the best fit to the items from SBS 2009 exam science subtest through various goodness of fit analysis. By means of goodness of fit analysis, both the IRT model assumptions and expected model features were investigated.

Investigations of IRT model assumptions indicated that after deleting five items, the unidimensionality assumption was met. As a result of this situation, it was assumed that the local item dependence assumption was also met.

The invariance of ability and item statistics obtained by each IRT model was also tested. The calculated correlations under each IRT model for ability and item parameters are strong enough. Therefore, it can be stated as the invariance property of ability and item parameters was held.

The analysis presented that 3-PLM provides the most appropriate fit to science subtest data. In other words the 3PLM provides the best item and ability parameter invariance which can be get after the model-data-fit is satisfied.

Finally, the deleted items may reduce the content validity of the subtest. Therefore, validity of the test should also be investigated and if necessary the analysis should be done via the multidimensional models.

References

Baker, F. B. (2001). *The basics of item response theory*. http://edres.org/irt/. adresinden alındı

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart & Winston.

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Hambleton, R. K. & Russell, W.J. (2005). *Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development.*An NCME Instructional Module

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* USA: Kluwer-Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory.* California: SAGE Publications.

McDonald, R. (1999). *Test Theory: A Unified treatment.* Mahvah, NJ: Lawrence Erlbaum Associates Publishers.